DOCUMENT RESUME

ED 069 762 TM 002 217

AUTHOR Baker, Eva L.
TITLE Using Measurement to Improve Instruction.
PUB DATE Sep 72
NOTE 8p.; Paper presented at annual meeting of American Psychological Association (Honolulu, Hawaii, Sept., 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Academic Achievement; *Achievement Tests; Criterion Referenced Tests; *Instructional Improvement; *Measurement Techniques; Norm Referenced Tests; Objective Tests; Psychometrics; Speeches; Student Evaluation; Teacher Role; *Test Construction; Testing; *Test Interpretation
IDENTIFIERS Domain Referenced Tests

ABSTRACT

Instructional improvement within the context of criterion-referenced and norm-referenced tests is described. Such categories overemphasize test interpretation rather than design characteristics of achievement tests. Data from most measurement situations may be reported or interpreted either according to criterion- or norm-referenced standards. How the test is developed and what it represents is of critical importance. The paper proposes alternative conceptualizations of test design: construct-referenced, objectives-referenced and domain-referenced. Using student data, the teacher needs to identify deficiencies in achievement, possible explanations, and remedies and to put the remedies into operation. An analysis of the utility of each test type results in the appraisal that domain referenced tests provide the most information for teachers and therefore are the most desirable as data sources for instructional improvement. However, because of lack of knowledge about instruction, poor training in available instructional principles, and lack of resources to encourage changes in instructional habits, it is concluded that instructional improvement, even if measurement considerations were satisfied, is not imminent. (Author/DJ)

ED 069762

TM 002 217

USING MEASUREMENT TO IMPROVE INSTRUCTION*

Eva L. Baker

University of California, Los Angeles

In uninventive fashion, I shall begin with a list of definitions, qualifications, caveats and platitudes to place later heresies in context.

First, the term "instruction." Assume that we mean the arrangement of conditions and events through which learning is presumably facilitated. For this discussion, accept the admittedly limited definition of instructional improvement in terms of pupil growth on some measure, rather than a refinement of a prescribed set of teacher behaviors.

Instruction can be mediated by a teacher, a set of materials, or some combination of the two. Based on a measurement, the teacher alters procedures in some way to effect better results. Similarly, the designers of materials will re-work them, or their support systems, to produce better pupil performance on a subsequent measurement. Instructional improvement is usually conceived in terms of the particular curriculum goals of the institution, for instance, a teacher's ability to bring about reading gains or the effectiveness of materials in teaching classification of concepts. Such instruction operates in a network of constraints. Happenstance, such as whether the children previously had a good arithmetic teacher (that is, did they learn arithmetic well?) plays an important role and limits the extent to which a teacher can determine or improve his or her instructional competence in teaching higher mathematical principles. Such curriculum-linked instruction is also naturally wedded to the available or approved instructional texts

*Paper presented at a Symposium, "The Relative Strengths of Norm-Referenced and Criterion-Referenced Achievement Tests," of the Annual Meeting of the American Psychological Association, Honolulu, September, 1972.

FILMED FROM BEST AVAILABLE COPY

and aids. If a school district provides insufficient numbers or inferior texts for students' use, it is likely that, for many teachers, instructional improvement is circumscribed.

Whether fully or less constrained by school limitations, instructional improvement needs, as a point of departure, a set of measurements. How can these measures be employed to improve the learning in the schools? The distinction between norm-referenced and criterion-referenced tests is not helpful to me, for the terms over-emphasize interpretation of the tests. More important is the basis for test construction and the instructional implications which flow from design, rather than test interpretation. Obviously, data from most measurement situations may be reported or interpreted by comparing the number of items obtained against the number of items available or with any other arbitrary standard; test data may also be reported by comparing a child or group's achievement level (whatever it was) to performance of other children. The critical factor in instruction is not how the results are portrayed, for that is a subsequent problem, but how they are obtained and what they presumably represent. If norm and criterion referenced tests are not appropriate descriptors to differentiate among the design characteristics of tests for use in instructional settings, perhaps other categories should be explored.

Instead of "norm-referenced" tests, I suggest construct-referenced to describe achievement tests which consist of a wide variety of item types and a relatively well-sampled content range. Such a label is intended to be independent of the manner in which the tests are ultimately interpreted, but could probably be applied to many present commercially produced and widely used achievement tests.

Labeling what passes for "criterion-referenced" tests is more difficult. The first alternate title is objective-referenced tests. However, such designation is unfortunately misleading, for it does not follow that, if one has an objective based on observable behavior, one will produce homogeneous test items which relate to the objective. In fact, since the content specifications are often poor, one can depend only on the fact that item formats of objectives-referenced tests will be similar, e.g., all short answer; all multiple choice with four options.

A substantial refinement over objectives-referenced tests are domain-referenced tests. (See Hively, et al., 1968, 1971.) Instead of a "behavioral" or performance objective emphasizing, for instance, that the learner will be able to pronounce phonemic combinations, a domain specifies both the performance the learner is expected to demonstrate as well as the content domain to which the performance is to generalize. In the pronunciation example, either the content of interest (sh, th) or a generation rule (all ending blends) for content is provided. Such tests attempt to clarify what it is they are attempting to measure and to provide a fuller basis for revision by a potential user.

To summarize, consider three different types of achievement tests for instructional improvement: construct-referenced, objective-referenced and domain-referenced. The emphasis in construct-referenced tests is on providing a full range of content and behaviors relevant to a construct such as computational ability. The emphasis of objectives-referenced tests has been on providing items which exhibit similar response requirements related to an often poorly defined content area, e.g., an objective which states the child will be able to write the theme of an essay when the critical properties of

essays is not explained. The domain-referenced tests includes items which conform to a particular response requirement, such as pronunciation, and provides a description as well as the class of content to which the performance is presumably to generalize, i.e., consonant-vowel-consonant words.

The three test types have political implications as well. Construct-referenced tests, by their published titles, promise grand things, for they measure areas like "critical reading," and "scientific concepts." Children who perform poorly in such measures are treated with head-shaking pity. Objective-referenced tests also contract for more than they deliver. A test which measures the child's ability to derive meaning from paragraphs by answering questions, will surely miss a range of paragraph and question complexity which critics feel is important. But because an objective has been written, it may appear to a user, such as a school board, that there is a great deal of specifity in the goal and thus someone (teachers) should know enough to achieve it. Tests which appear precise but are not can seriously mislead teachers and administrators.

Domain-referenced tests have not been developed frequently enough to promote predictable responses in users. However, since their content is so well defined one would expect a fuller congruence with the user's needs and the test's purposes. Domain-referenced tests are so time-consuming to produce that only a relatively few will ever be satisfactorily written, and those only for critical, consensus objectives.

If teachers had useful information from tests, so the story goes, instructional improvement would follow. Approaches to instruction, characterized by their proponents as "decision-oriented," "competency-based," "rational," or

"systematic" are centered on the premise that if teachers could be provided with valid information on the performance of their students, they would be able to adapt their instruction and successfully remediate.

Posit a minimum set of events and knowledge that a teacher needs in order to implement an instructional improvement cycle:

Step 1. Data on students' abilities to perform skills and behaviors.

Step 2. Ability to identify deficiencies in students' achievement.

Step 3. Ability to identify possible explanations for these deficiencies.

Step 4. Ability to identify alternative remedial sequences.

Step 5. Ability to implement such sequences.

Such an event set requires, at minimum, compliance by students, something frequently not guaranteed. Compare the three types of tests, construct, objectives and domain-referenced in terms of how they might facilitate the instructional cycle. All three tests provide useful (Step 1) data. Construct-referenced tests are presently most respectably developed. They are, however, administered on a schedule not normally consistent with continuing diagnosis, and are often reported in terms of the childs' status with respect to other children rather than his or her own particular strengths and weaknesses. Still, a teacher could get a general idea about learners' proficiency. Objectives-referenced tests may be scheduled more regularly and provide data which appear to give information about what the child can do, but because the content analysis in the test design is usually weak, these tests may not provide serious assistance in helping a teacher to identify actual competencies. Domain-referenced tests represent an improvement in the quality of information they provide, in that the range of instances to which a learner is able to perform is explicit. Data from such tests are "enabling;" if teachers would,

they could identify with increased explicitness what the students were able to deal with. Identification of performance deficiencies (Step 2), is theoretically possible through the use of all three tests. Since arbitrary judgments are usually invoked in deciding on what constitutes a deficiency, that is, the 44th percentile is bad, or 68 percent is unsatisfactory, none of the test types seriously advantages the user. Deficiency, even if there were defensible procedures for determining cut-off points to define "deficiency," we would halt the analysis of the utility of measurement to foster instructional improvement.

Even if test producers get very busy and produce a range of exciting, important and valid achievement instruments, many teachers would be unable to put the data produced to reasonable use (Steps 3, 4, 5) for the following reasons. First, only limited knowledge is available in the instructional field. Even agencies with talented instructional designers treat each development task, in large measure, as wholly idiosyncratic and employ heuristic test-and-revision cycles in the validation of materials. Well-researched instructional principles exist in only limited, and largely operant, clusters.

Secondly, even where instructional design principles exist, they are not disseminated. Although many teachers function well without arcane knowledge from instructional research, less gifted teachers might be able to put such knowledge to use, but they have no access to the fount. Teaching training, in disarray for years, has not yet provided adequate preparation for many. Coordinators of in-service education of teachers rarely have sufficient resources to provide training. When well-funded, expertly staffed instructional development agencies spend considerable time designing and redesigning satisfactory instruction in one or two areas, why should one expect a single teacher, modestly trained, to be able to do as well in many subject matters with few resources?

Beyond the paucity of instructional principles and the dearth of training is the nature of the individual teacher's predispositions. Even if: (1) good data were available, (2) reasonable "criterion" levels were agreed to, (3) instructional principles existed, (4) teachers know how to use such principles and adapt them to given situations, habitual instructional routines will need to be overcome. Teachers will need incentives, support, and rewards if they are to change significantly their present practices. In fact, since most accountability systems use the threat of punishment rather than incentives as a basis for fostering teaching improvement, one could become even more pessimistic about the likelihood of facilitating teacher change.

If analysis leads one to believe that, even with measurement advances, instructional improvement would not inevitably follow, what implications are there for research and development activity in test design? Further, in the present accountability surge what can instructional and measurement experts do to help both the teachers and the students? Clearly, construct-type tests will continue to be used to give a broad, comparative picture of school achievement and they should be. Objectives-referenced tests may be appropriate for individual teachers to use to measure their pupil's progress and their own achievement of certain goals of high personal interest. They should probably be locally prepared, since technical quality of the tests will necessarily relax, and results should be of interest on a personal classroom-feedback level only. Domain-referenced tests are those tests which may be employed for evaluation, as in accountability, where improvement is expected. The use of domain-referenced data, gives the teachers most

assistance, for they are provided with clear information about what kind of practice items are in the set of content and performance measured by the test. One might expect that teachers could be easily prepared to provide instructional situations that allow students to practice content from the appropriate set without permitting the students to have experience with the test items themselves. Domain-referenced tests are difficult to prepare, particularly because not all subject matter is presently analysed in a way to permit the preparation of such tests. If experts in American government insist that there are in fact three functions of the executive branch in the United States, then no amount of analysis by skilled psychometricians to come to deeper truths is worthwhile. Where subject matter experts cannot provide appropriate and generalizable dimensions for the analysis of subject matter, psychometricians should not bear the burden of the trivia. It is not their problem. Perhaps, after all, such objectives should not be measured in any organized or institutional sense. I would suggest that relatively few areas be identified for accountability-teacher improvement testing. Basic reading and arithmetic speed into focus here. Beyond those two areas, I would suggest domain-referenced or objective based measurement be <u>publicly</u> used very sparingly. Other process-type measures could tell the taxpayer if the teachers all performing adequately, until teachers are trained and willing to use appropriate instructional strategies, the quest for valid achievement measurements will remain a challenging problem, but one functionally irrelevant to arena of instructional improvement.

BJ4